JOURNAL OF BACTERIOLOGY, Mar. 2004, p. 1518–1530 Vol. 186, No. 5
0021-9193/04/$08.00+0 DOI: 10.1128/JB.186.5.1518–1530.2004

eBURST: Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequence Typing Data

Edward J. Feil,[1]* Bao C. Li,[2] David M. Aanensen,[2] William P. Hanage,[2] and Brian G. Spratt[2]

Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY,[1] and Department of Infectious Disease Epidemiology, Imperial College London, St. Mary's Hospital Campus, London W2 1PG,[2] United Kingdom

Received 29 August 2003/Accepted 14 November 2003

The introduction of multilocus sequence typing (MLST) for the precise characterization of isolates of bacterial pathogens has had a marked impact on both routine epidemiological surveillance and microbial population biology. In both fields, a key prerequisite for exploiting this resource is the ability to discern the relatedness and patterns of evolutionary descent among isolates with similar genotypes. Traditional clustering techniques, such as dendrograms, provide a very poor representation of recent evolutionary events, as they attempt to reconstruct relationships in the absence of a realistic model of the way in which bacterial clones emerge and diversify to form clonal complexes. An increasingly popular approach, called BURST, has been used as an alternative, but present implementations are unable to cope with very large data sets and offer crude graphical outputs. Here we present a new implementation of this algorithm, eBURST, which divides an MLST data set of any size into groups of related isolates and clonal complexes, predicts the founding (ancestral) genotype of each clonal complex, and computes the bootstrap support for the assignment. The most parsimonious patterns of descent of all isolates in each clonal complex from the predicted founder(s) are then displayed. The advantages of eBURST for exploring patterns of evolutionary descent are demonstrated with a number of examples, including the simple Spain^{23F}-1 clonal complex of *Streptococcus pneumoniae*, "population snapshots" of the entire *S. pneumoniae* and *Staphylococcus aureus* MLST databases, and the more complicated clonal complexes observed for *Campylobacter jejuni* and *Neisseria meningitidis*.

The ability to accurately determine the genetic relatedness of isolates of bacterial pathogens (or other disease agents) is fundamental to molecular epidemiological and evolutionary studies. In recent years, the use of nucleotide sequence variation at multiple housekeeping loci has become increasingly popular for strain characterization, as it has advantages for inferring levels of relatedness between strains and the reconstruction of evolutionary events (1, 2, 6–14, 18–23, 25, 28, 29).

In many bacterial species, genetic variation at housekeeping loci accumulates as frequently or more frequently by homologous recombination (replacement of small chromosomal segments with those from related isolates) as by point mutation (15). Over the long term, recombination may prevent the true relationships between distantly related isolates of a species from being discerned. Epidemiological studies, however, are typically concerned with disease outbreaks or the spread of antibiotic-resistant or virulent strains between countries. Over these very short evolutionary timescales, of weeks to a few hundred years, recombination is unlikely to prevent the recognition of clones and clonal complexes within most bacterial populations. Thus, although the phylogenetic complexities introduced by homologous recombination may be problematic over long periods of evolutionary time (14, 15), given an appropriate model of bacterial evolution, it should be possible to accurately reconstruct evolutionary events that occur over short timescales, even if rates of recombination are high.

Characterization of isolates of bacterial pathogens on the basis of sequence variation is carried out by multilocus sequence typing (MLST), which generates approximately 450 bp of nucleotide sequence for internal fragments of seven housekeeping loci for each isolate (23, 33). The different sequences at each locus are assigned different allele numbers, and each strain is defined by the alleles at the seven loci (the allelic profile). Each unique allelic profile (or genotype) is assigned a sequence type (ST), which is a convenient and unambiguous descriptor for the strain (or clone). Analyses of isolates of several bacterial species by MLST (7–10, 23, 25) support the view obtained from earlier studies (3, 26, 32) that a considerable proportion of a population belongs to a limited number of clusters of closely related genotypes, here referred to as clonal complexes. Clonal complexes are typically composed of a single predominant genotype with a number of much less common close relatives of this genotype (15).

The simplest model for the emergence of clonal complexes is that a founding genotype increases in frequency in the population, as a consequence either of a fitness advantage or of random genetic drift, to become a predominant clone (15). As it increases in frequency in the population, the founding genotype gradually diversifies, to result in a clonal complex. In terms of MLST, descendants of the founder will initially remain unchanged in allelic profile, but over time variants in which one of the seven alleles has changed (by point mutation or recombination) will arise. These genotypes, which have allelic profiles that differ from that of the founder at only one of

* Corresponding author. Mailing address: Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom. Phone: 44 1225 383021. Fax: 44 1225 386779. E-mail: e.feil@bath.ac.uk.

the seven MLST loci, are called single-locus variants (SLVs). Eventually, SLVs will diversify further, to produce variants that differ at two of the seven loci (double-locus variants [DLVs]), at three of the loci (triple-locus variants [TLVs]), and so on.

MLST data are typically represented by a dendrogram (e.g., the unweighted pair-group method with arithmetic averages [UPGMA]) on the basis of a matrix of pairwise differences in the allelic profiles of the isolates. This dendrogram provides a convenient means of identifying isolates that are identical or closely related in genotype and that can be assigned to the same clone or clonal complex. However, the topology of such dendrograms can be somewhat arbitrary, and they provide essentially no information on the patterns of evolutionary descent of the isolates within a clonal complex or the identity of the founder.

Here we describe a new implementation of the BURST algorithm, called eBURST. This approach subdivides large MLST data sets into nonoverlapping groups of related STs or clonal complexes and then discerns the most parsimonious patterns of descent of isolates within each clonal complex from the predicted founder. As this approach is dependent upon the correct assignments of founding genotypes, a bootstrapping procedure is introduced to gauge the level of confidence in these assignments. Through a very simple set of rules, eBURST can be used to explore how bacterial clones diversify and can provide evidence concerning the emergence of clones of particular clinical relevance. We demonstrate the utility of this approach by using MLST data from antibiotic-resistant *Streptococcus pneumoniae* (8, 35) and from *Staphylococcus aureus* (9, 11, 16), *Campylobacter jejuni* (7, 31, 34), and *Neisseria meningitidis* (20, 23). The rapid rate of clonal diversification in the latter two species (13, 31) provides a particularly challenging test of procedures that aim to untangle the short-term evolutionary history of a bacterial species.

MATERIALS AND METHODS

eBURST algorithm. The eBURST algorithm is implemented as a Java applet at http://eburst.mlst.net, and detailed guidance in its use is available at this website. A description of the algorithm is given below.

Subdivision of input data into groups. The first step of the eBURST algorithm is to subdivide STs into groups. Every ST within an eBURST group has a user-defined minimum number of identical alleles in common with at least one other ST in the group. eBURST groups therefore are mutually exclusive; no ST can belong to more than one group. The default setting in eBURST is the most exclusive group definition, in which STs are included within the same group only if they share identical alleles at six or seven of the seven MLST loci with at least one other ST in the group. Thus defined, each group equates to a single clonal complex (see below).

The above procedure provides a list of the STs assigned to each group along with their observed frequencies in the data set. STs that cannot be assigned to any group are called singletons. For example, with the default group definition, singletons are defined as STs differing at two or more alleles from every other ST in the sample. eBURST also allows all of the input data to be treated as a single group by selecting a group definition of zero of seven shared alleles. This procedure allows the clustering patterns among all isolates within a complete MLST database to be visualized as a single eBURST diagram ("population snapshot").

eBURST groups and clonal complexes. We draw an important distinction between an eBURST group and a clonal complex. Whereas an eBURST group is simply a collection of STs that are placed together according to the selected group definition, a clonal complex refers to a biologically meaningful cluster of STs that have diversified very recently from a common founder. The STs within an eBURST group obtained with the most stringent (exclusive) group definition are closely related and are considered to belong to a single clonal complex. Groups obtained with a less stringent group definition should not be equated with clonal complexes. eBURST displays only the most likely patterns of evolutionary descent within each clonal complex and does not attempt to reconstruct pathways between clonal complexes, even if they are closely related. Clonal complexes therefore are defined conservatively as a cluster of STs in an eBURST diagram in which all STs are linked as SLVs to at least one other ST. These may be represented individually when the default group definition is used or as separate clusters of linked STs when a less stringent group definition is used.

Assignment of primary founders. For each clonal complex, eBURST identifies the ST that is most likely to represent the founding genotype (the primary founder). eBURST also attempts to identify the most likely founder of a group when a more relaxed group definition is used, although often in such situations the assigned founder is unlikely to represent the original genotype of the entire group. The primary founder is predicted on the basis of parsimony as the ST that has the largest number of SLVs in the group or clonal complex. This method of assigning the founder takes into account the way in which clones emerge and diversify; most of the initial diversification of a clone results in variants of the founder that differ at only one of the seven alleles (i.e., SLVs of the founder). If two STs in a group have the same number of SLVs, then the one with the larger number of DLVs is chosen. In such situations, the confidence in the assignment is low, as reflected in the bootstrap values. In some groups, typically those composed of a very limited number of STs, it may not be possible to assign a primary founder. The frequency of a given ST in the input data is not used in the procedure to assign founders; however, founders often correspond to the most predominant STs, a fact that adds independent support to the assignments.

Assigning levels of confidence in founding genotypes. A measure of statistical confidence in each of the assigned primary founders is made by a bootstrap resampling procedure. eBURST is used to divide the input population into groups according to the selected group definition; for each group, one example of each ST is extracted, and a user-defined (default, 1,000) number of random data sets of the same size as the extracted ST set is produced by resampling with replacement. eBURST is run on the resampled data sets from each group, and the ST that is assigned as the primary founder in each resampling is determined. Conditional bootstrap values for each ST in the group are generated according to the percentage of times that the ST is assigned as the founder; resamplings in which the ST cannot be assigned as the founder due to its absence from the resampled data set are omitted. An ST that is assigned as the founder in each of the resamplings in which it is present therefore has a bootstrap value of 100%. The computation of bootstrap values is restricted to gauging the confidence of the assignment of primary founders for individual clonal complexes by using the default (most stringent) group definition.

Assignment of subgroups and subgroup founders. Large clonal complexes typically contain subgroups and therefore have both primary and subgroup founders. For example, an SLV of the primary founder may have increased in frequency and diversified to generate a number of its own SLVs, thus becoming a subgroup founder. The promotion of an ST to a subgroup founder depends upon the number of previously unassigned SLVs that it defines (see below). This definition can be user defined, but the default setting is at least two previously unassigned STs (i.e., at least three links to other STs, including the link to its assumed progenitor).

A single ST may be an SLV of more than one founder. When an ST is an SLV of both primary and subgroup founders, the ST is preferentially assigned to the primary founder. When an ST is an SLV of two or more subgroup founders, eBURST initially assigns SLVs on the basis of the distance from the primary founder, but a local optimization procedure then reassigns SLVs preferentially to the largest subgroup. In this way, the same model of clonal expansion is used for both primary and subgroup founders while links between subgroups are retained (see the readme file at http://eburst.mlst.net for more details). Although any given ST may be an SLV of more than one founder, this procedure allows an ST to be assigned to only one founder. This means that the number of SLVs of each subgroup founder shown in the eBURST diagram often is smaller than the total number shown in the initial eBURST output table, as some SLVs will have been preferentially assigned to the primary founder or to other subgroup founders.

Text and graphical output from eBURST. eBURST has a variety of input options (including direct input from MLST databases; http://www.mlst.net) and produces an output table defining the STs in each group, the number of isolates of each ST, and the number of SLVs, DLVs, and TLVs of each ST. The predicted primary founder is identified (where possible) along with the percent recovery of each ST as the primary founder of the clonal complex in the bootstrap resamplings.

The eBURST diagrams display the patterns of descent of all STs within each clonal complex from the primary founder. The earlier version of the algorithm (BURST) positioned SLVs and DLVs of the primary founder within concentric

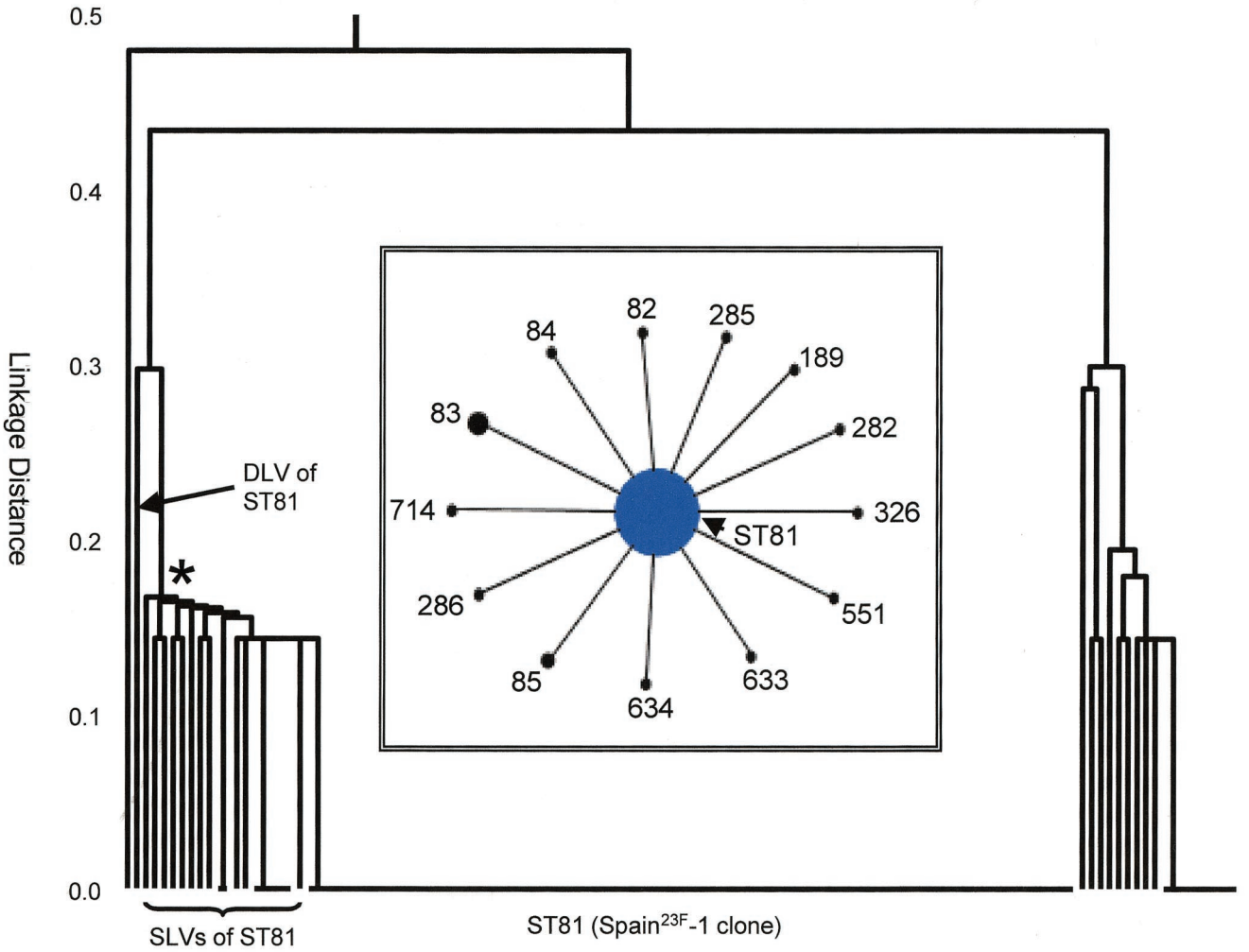

FIG. 1. Analysis of the ST81 clonal complex of *S. pneumoniae*. The relatedness between isolates in the pneumococcal MLST database that shared alleles at four or more loci with the allelic profile of ST81 (Spain^{23F}-1 clone) is displayed as a dendrogram. The entire pneumococcal MLST database was analyzed by eBURST with the stringent (default) group definition; the group that included ST81 is displayed as an eBURST diagram (inset). Numbers in the eBURST diagram correspond to ST numbers. The STs in the eBURST diagram included all of those arising from the node on the dendrogram marked by an asterisk. One DLV of ST81 (arrow) was not included in the eBURST group when the stringent group definition was used. The area of each circle in the eBURST diagram corresponds to the abundance of the isolates of the ST in the input data; ST81 is the predicted founder of the group (bootstrap confidence value of 100%).

rings (11, 15), whereas eBURST shows a radial link from the primary founder to each of its SLVs by a solid line. A second difference is that only links to SLVs are shown; DLVs of the primary founder are linked only when the intermediate SLV on the path from the founder to the DLV is present in the input data. With the default group definition (six of seven shared alleles), all STs must be SLVs of at least one ST in the group, and the eBURST diagram will show a single cluster (a clonal complex) in which all STs are linked. With the less stringent group definition (five of seven shared alleles), more than one cluster of linked STs (each of which is a clonal complex) may be displayed along with a number of individual unlinked STs. The lack of linking between two clusters within a single group implies that no ST in one cluster is an SLV of any ST in the other cluster. Similarly, individual unlinked STs are not SLVs of any ST in the group. Thus, eBURST is very conservative and only shows links between STs that have diverged very recently, that differ at only a single locus, and that are considered to belong to the same clonal complex.

Each ST is represented as a circle; the number beside the circle is the ST (except in Fig. 1, the ST numbers have been removed for increased clarity). The frequency of each ST (i.e., the number of isolates of the ST in the input data) is indicated by the area of the circle. The primary founder is given in blue, while subgroup founders are given in yellow. The initial eBURST diagrams were edited, as required, to produce the final figures; details of the editing functions within eBURST are provided at http://eburst.mlst.net. Editing only changes the positions of the STs to improve the clarity of the diagram and does not change any of the links between the STs.

Input data. The complete current sets of isolates (as of July 2003), with their STs and allelic profiles, were extracted from public MLST databases at the following websites: *S. pneumoniae* (http://spneumoniae.mlst.net; 1,638 isolates, 893 STs); *S. aureus* (http://saureus.mlst.net; 1,072 isolates, 191 STs); *C. jejuni* (http://campylobacter.mlst.net; 2,001 isolates, 796 STs); and *N. meningitidis* (http://neisseria.mlst.net; 3,730 isolates, 2,609 STs). eBURST was applied to the entire database for each species with the group definitions specified to identify the groups, and eBURST diagrams were generated.

Construction of trees. UPGMA dendrograms were constructed from the matrix of pairwise differences in the allelic profiles of the isolates by using the Statistica package (Statsoft Inc., Tulsa, Okla.).

RESULTS

A multiply antibiotic-resistant clone of *S. pneumoniae*. As discussed above, eBURST is primarily an epidemiological tool designed for examining clonal diversification over short evolu-

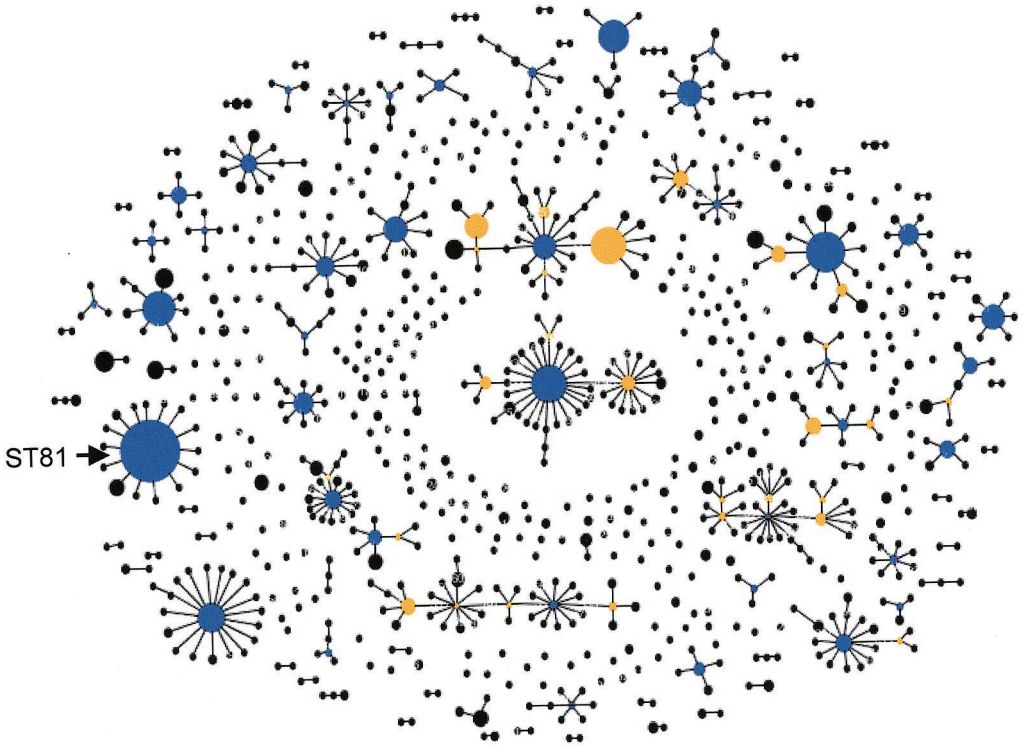

FIG. 2. Population snapshot of *S. pneumoniae*. Clusters of related STs and individual unlinked STs within the entire pneumococcal MLST database are displayed as a single eBURST diagram by setting the group definition to zero of seven shared alleles. Clusters of linked isolates correspond to clonal complexes. Primary founders (blue) are positioned centrally in the cluster, and subgroup founders are shown in yellow. Only the ST81 cluster shown in Fig. 1 is labeled; the other ST labels have been removed for clarity.

tionary timescales. Antibiotic-resistant strains therefore provide a simple test case, as these are unlikely to predate the introduction into medicine of the antibiotics to which they show resistance and should have diversified little from their primary founder within this very short period of time.

Strains of *S. pneumoniae* that are resistant to multiple classes of antibiotics were first reported from South Africa and Spain in the late 1980s; one of the first of these to be characterized is the Spanish multiply antibiotic-resistant serotype 23F clone (Spain^{23F}-1) (24). The majority of isolates assigned as Spain^{23F}-1 by molecular typing methods have been shown by MLST to have the same allelic profile (ST81) (35). All isolates with an allelic profile similar to that of ST81, sharing four or more of the seven MLST alleles, were extracted from the pneumococcal MLST database. In order to compare the results from eBURST with those from more traditional techniques, a UPGMA dendrogram was constructed from the matrix of pairwise differences in the allelic profiles of the extracted isolates. Figure 1 shows multiple isolates of ST81, a cluster of STs very closely related to ST81 (all SLVs of ST81), and one slightly more distantly related ST (a DLV of ST81); all of these isolates were multiply antibiotic resistant. None of the isolates on the dendrogram that were more distantly related to ST81 (linkage distance of greater than 0.4) were multiply antibiotic resistant, and they were very unlikely to have descended from ST81.

The entire pneumococcal MLST database was entered into eBURST, and groups were defined with the stringent (default) group definition (six or more shared alleles). The group containing ST81 was displayed as an eBURST diagram (Fig. 1, inset). Consistent with its high frequency, ST81 was assigned as the primary founder of the ST81 (Spain^{23F}-1) clonal complex, with bootstrap support of 100%; all of the other isolates in this eBURST group were SLVs of ST81. The prevalence of ST81 in the input data set is reflected by the area of the circle in the eBURST diagram. The one DLV of ST81 was not included, as the linking SLV was not present in the MLST database, although it was included when the group definition was made less stringent (five of seven shared alleles). The structure of this clonal complex therefore is simple, with the founder radially linked to its 13 SLVs, reflecting the very short evolutionary timescale over which ST81 has diverged (less than 50 years).

Pneumococcal population snapshot. The ability of eBURST to provide an overview of the clonal complexes within an entire MLST database was demonstrated by an analysis of all 1,638 isolates in the pneumococcal MLST database, accounting for 893 STs. All isolates were analyzed as a single group by setting the group definition to zero of seven shared alleles; the eBURST diagram is shown in Fig. 2. The diagram shows the major clusters of linked STs (clonal complexes), the minor clusters, linked triplets and doublets, and individual unlinked STs. The ST81 clonal complex shown in Fig. 1 is labeled. Note that the spacing between unlinked STs and clonal complexes provides no information concerning the genetic distance between them.

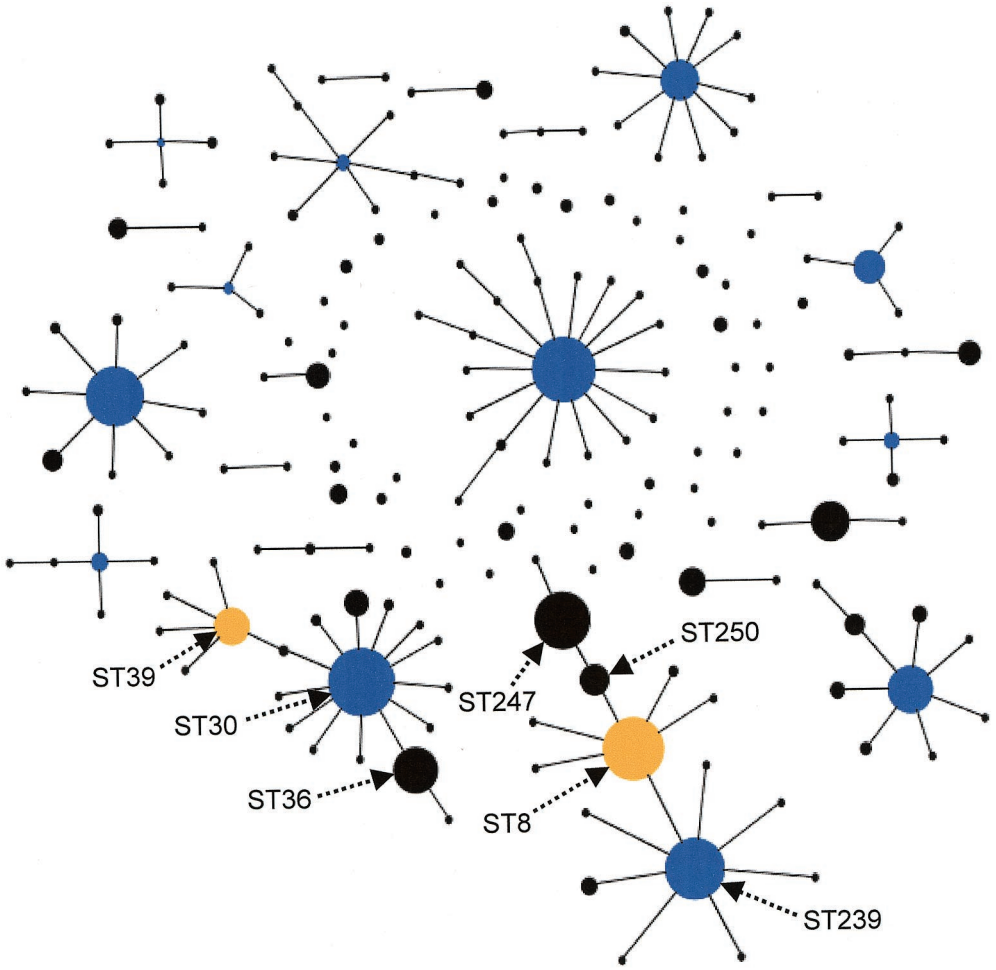

FIG. 3. Population snapshot of *S. aureus*. The entire *S. aureus* MLST database is displayed as a single eBURST diagram as described in the legend to Fig. 2. The major STs within the ST30 and ST239 clonal complexes are marked by arrows; the patterns of descent within these complexes are discussed in the text. For clarity, ST labels have been removed.

Evolution of MRSA. *S. aureus* is an important gram-positive human pathogen and, since the early 1960s, methicillin-resistant *S. aureus* (MRSA) isolates have emerged. MRSA isolates are now particularly common in hospitals, although their prevalence is also increasing in the community. The gene conferring resistance to methicillin is transmitted horizontally through the *S. aureus* population, and MRSA clones are known to have emerged independently on multiple occasions (11, 27).

An extensive MLST data set is available for global collections of MRSA isolates, and the BURST algorithm was previously used to determine the origins of MRSA clones from their antibiotic-sensitive forebears (11). The MLST database for *S. aureus* as of July 2003 contains 1,072 isolates (191 STs) from global sources. These isolates are a mixture of MRSA and methicillin-susceptible *S. aureus* (MSSA) from disease cases and asymptomatic carriage. The eBURST diagram shown in Fig. 3 is the population snapshot of the entire *S. aureus* database showing the linked clusters of STs (clonal complexes), with the primary founders and subgroup founders identified. There were 12 clusters of four or more STs for which the primary founder could be assigned; many of these clonal complexes were previously described from a study of 334 isolates recovered from Oxfordshire, United Kingdom (9, 16). Interspersed among these clonal complexes were minor groups, typically doublets joined by an SLV link, and individual unlinked STs that were not SLVs of any other STs in the database.

Most of the clonal complexes of *S. aureus* are simple, with a primary founder surrounded by SLVs and, in some cases, DLVs. The ST30 and ST239 clonal complexes are more complicated (Fig. 3). The major ST30 clonal complex contains both MRSA and MSSA isolates (9, 11). ST30 is the predicted primary founder (99% bootstrap support) of this clonal complex. All of the MRSA isolates within this complex belong to ST36, with the exception of the one SLV of ST36 that has probably descended from it. ST36 is a well-characterized epidemic MRSA clone (EMRSA16) that appears to have emerged following the acquisition of methicillin resistance by an SLV of ST30 (11). One of the SLVs of ST30 appears to have diversified further to produce a DLV (ST39) that has become successful and that has formed a subgroup with its own SLVs.

The ST239 clonal complex includes the earliest known MRSA clone (ST250) (6, 11) and three other STs that represent MRSA clones commonly encountered within hospitals (ST8, ST239, and ST247) (6, 11). All three of these major STs

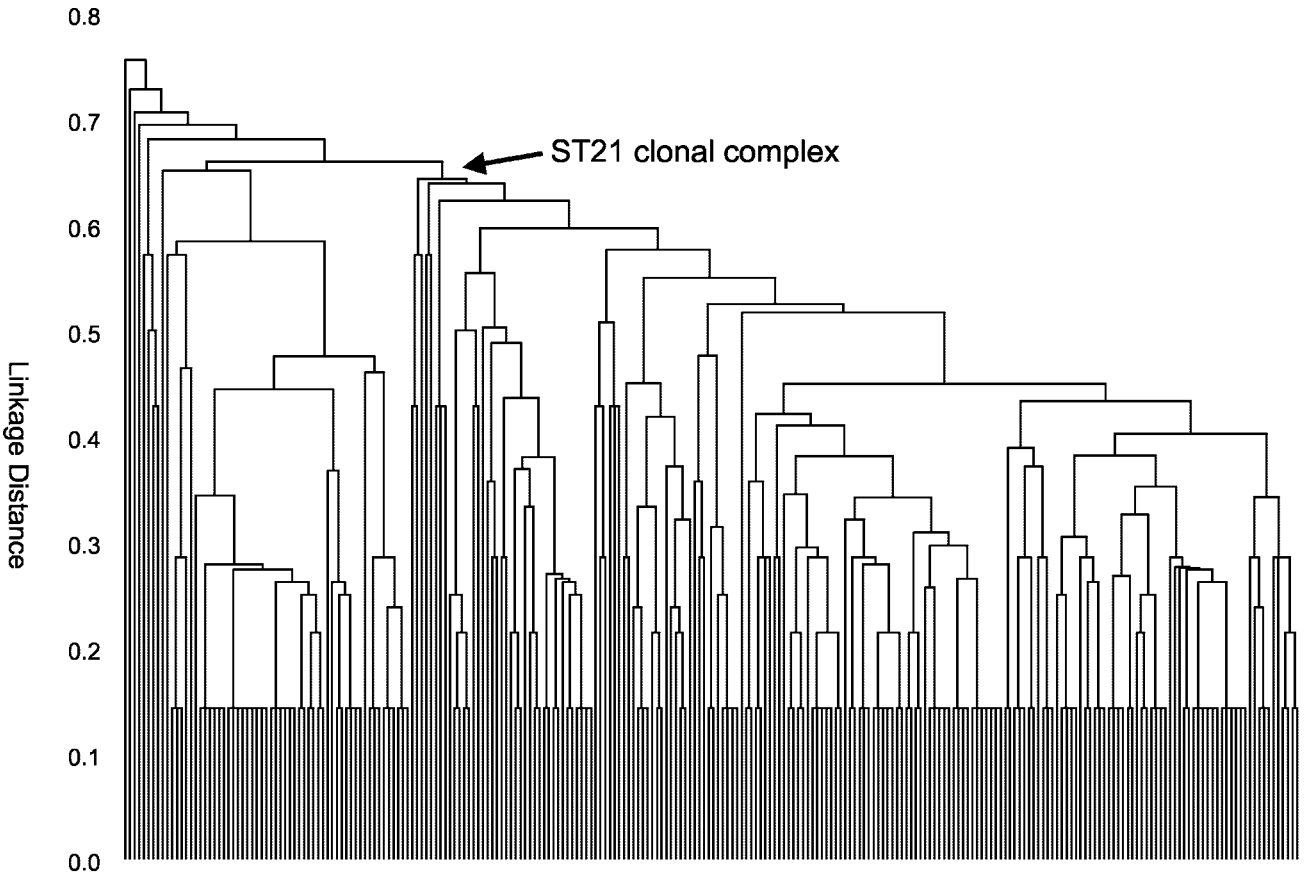

FIG. 4. Relationships of isolates of the *C. jejuni* ST21 clonal complex. STs that shared alleles at ≥3 of the 7 MLST loci with ST21 were obtained from the *C. jejuni* MLST website, and a dendrogram was constructed by using UPGMA. The node that defines the ST21 clonal complex is labeled on the dendrogram. Only one example of each ST is shown.

have diversified to produce their own SLVs, but as ST239 has the largest number of SLVs, it has been assigned as the primary founder of the clonal complex. This assignment is based on a single SLV; ST239 has eight SLVs, whereas ST8 has seven (one of these SLVs is preferentially assigned to the primary founder, as discussed in Materials and Methods). However, additional evidence, discussed by Enright et al. (11), suggests that ST8 rather than ST239 is the true founder of this clonal complex. The ambiguity in the assignment of the primary founder is reflected in the bootstrap support values obtained when this clonal complex is analyzed separately by eBURST with the default group definition of six of seven shared alleles. The bootstrap support values for ST239 and ST8 are 70% and 66%, respectively, thus alerting the user to the fact that the assignment of ST239 as the primary founder is not robust.

***C. jejuni* ST21 clonal complex.** *C. jejuni* is a gram-negative bacterial pathogen that causes gastroenteritis in humans and that is commonly isolated from chicken and cattle. An MLST scheme for this species was presented by Dingle et al. (7). The *C. jejuni* MLST database contains 2,001 isolates (796 STs). Recombination, which is believed to be frequent in this species (31), may lead to clones that diversify rapidly to produce complicated clonal complexes. The ST21 clonal complex is the largest within the *C. jejuni* database (7). The likely primary founder of this complex was identified by Dingle et al. (7) by using a combination of BURST and splits decomposition analysis. However, these authors did not use BURST to attempt to reconstruct the evolutionary pathways within this complex and instead used splits decomposition analysis for this purpose (7). Although this analysis confirmed ST21 as the most likely primary founder, the relationships between the STs were characterized by an extensive network, and recent patterns of descent could not be inferred.

Figure 4 shows a UPGMA clustering dendrogram containing one example of each ST that shares three or more alleles in common with ST21. This figure illustrates the size and complexity of the ST21 complex and the difficulties in inferring the most likely evolutionary pathways. The 2,001 isolates in the public *C. jejuni* MLST database were entered into eBURST and, with the stringent (default) group definition, the group including ST21 was identified. Figure 5 shows the eBURST diagram for the STs assigned to this clonal complex (688 isolates; 180 STs); many of these STs have been added to the public MLST database subsequent to the analysis by Dingle et al. (7). The analysis is consistent with that reported by Dingle et al. (7), in that ST21 remains the most likely primary founder (with 99% bootstrap support). This ST has 37 SLVs, whereas the two next most prevalent STs each have 24 SLVs. Several SLVs (and two TLVs) of ST21 (shown in yellow in Fig. 5) have

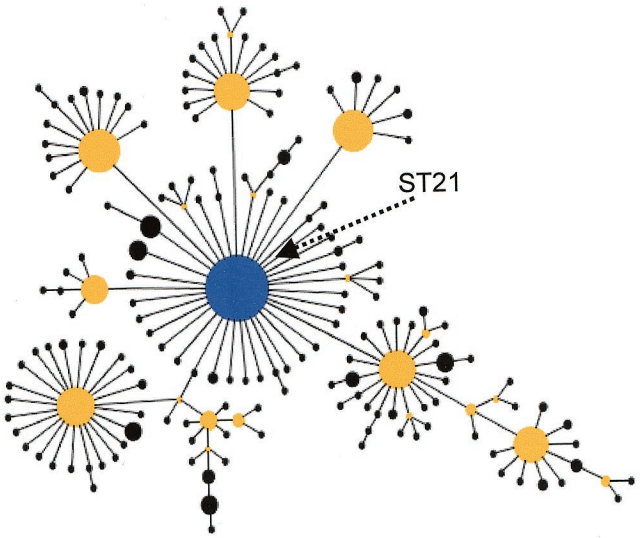

FIG. 5. Analysis of the ST21 complex of *C. jejuni*. The 2,001 isolates in the entire *C. jejuni* public MLST database were analyzed by eBURST with the stringent (default) group definition; the group that included ST21 is displayed as an eBURST diagram. The predicted primary founder, ST21 (bootstrap confidence value of 99%), is labeled.

emerged as successful subgroup founders, each with its own cluster of linked SLVs.

The primary and subgroup founders correspond to the STs that are the most prevalent within the ST21 complex. For example, ST21 is the most common ST within the group (123 isolates). Subgroup founders also are relatively common and, for the nine most common STs, there is a close relationship between the frequency of the ST and the number of SLVs of that genotype (data not shown).

***N. meningitidis* clonal complexes.** High rates of recombination also are a feature of the meningococcal population, and clones diversify rapidly (13). The entire public meningococcal MLST database (3,730 isolates; 2609 STs) was analyzed by eBURST with the stringent (default) group definition of six of seven shared alleles. Groups corresponding to the ET37 (ST11), A4 (ST8), and ET5 (ST32) clonal complexes (3, 4, 20, 23) were displayed as eBURST diagrams. Figure 6 shows the eBURST diagram for the ST32 clonal complex (4), which has a primary founder (ST32; 100% bootstrap support) surrounded by a ring of SLVs, one of which (ST33) has diversified to become a large subgroup founder. In addition, there are numerous DLVs and TLVs of the primary founder and the major subgroup founder.

In both the ST8 and the ST11 clonal complexes, there was also a single strongly supported primary founder (100% boot-

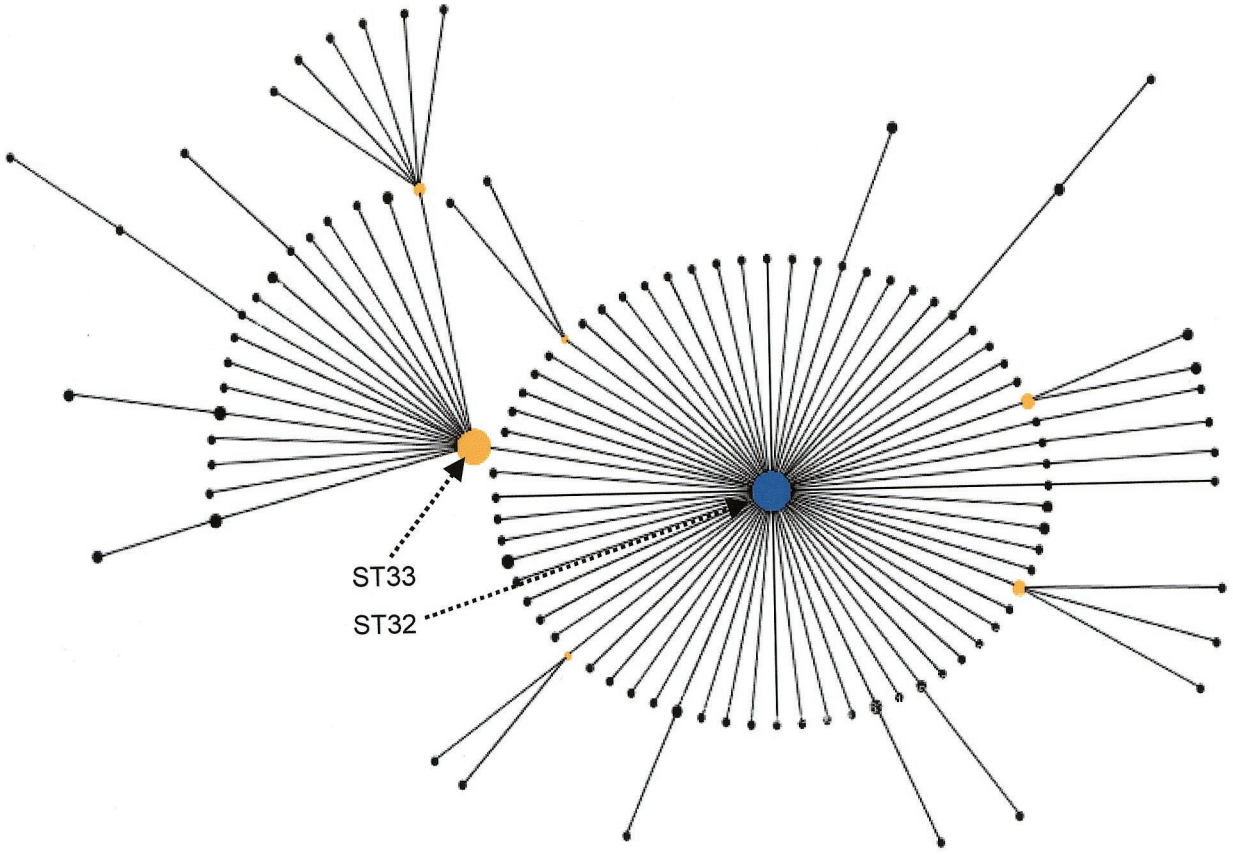

FIG. 6. Analysis of the ST32 clonal complex of *N. meningitidis*. eBURST groups were obtained from the entire meningococcal public MLST database with the stringent (default) group definition; the eBURST group that included ST32 is displayed. The primary founder, ST32 (bootstrap confidence value of 100%), and a major subgroup founder, ST33, are labeled.

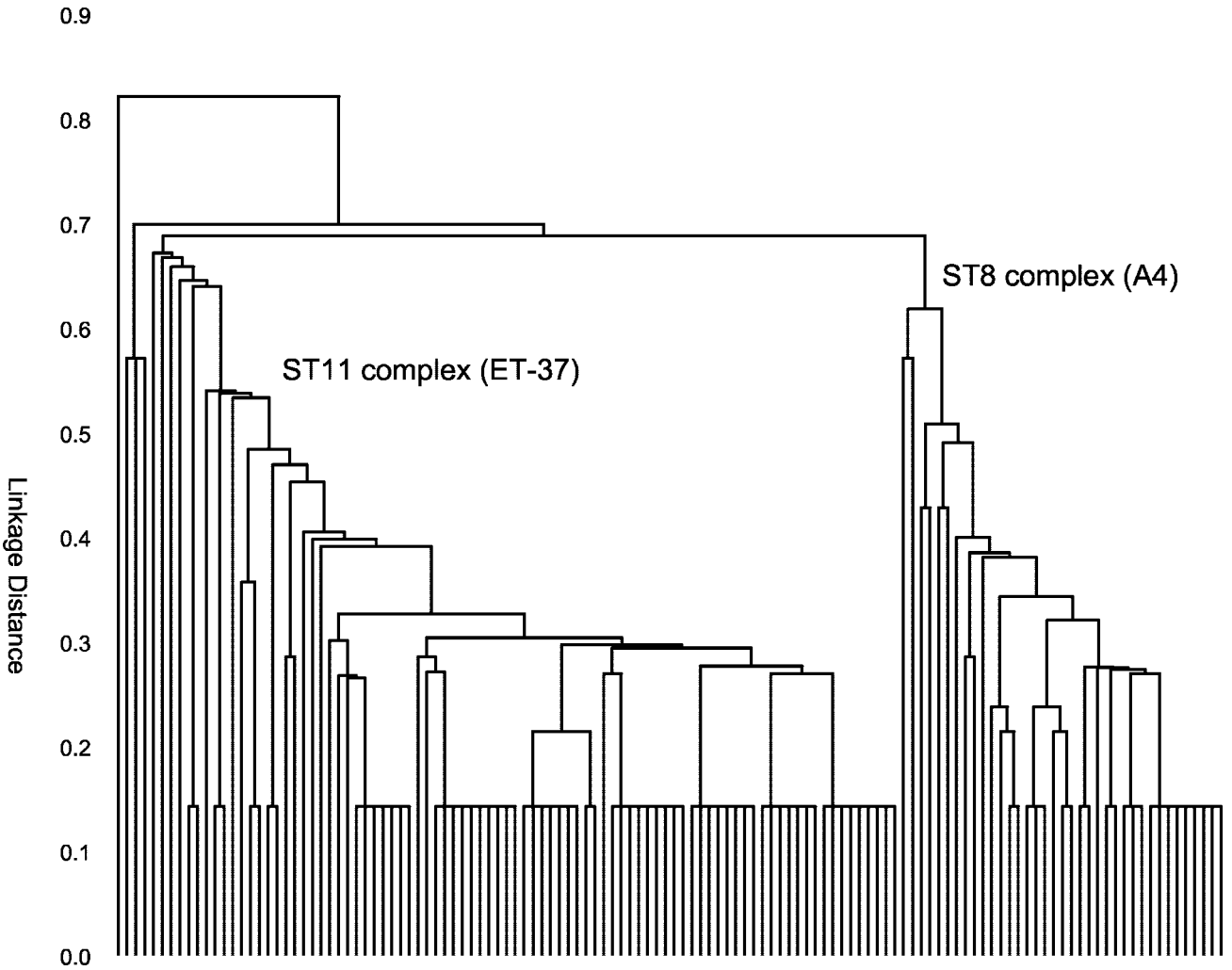

FIG. 7. Relatedness of STs of the ST8 and ST11 clonal complexes. STs that shared alleles at ≥3 of the 7 MLST loci with ST8 or ST11 were obtained from the *Neisseria* MLST website, and a dendrogram was constructed. The clusters of STs corresponding to the ST8 (A4) and ST11 (ET-37) complexes are shown. Only one example of each ST was used in the analysis.

strap values) and a simple pattern of diversification from the founder to produce a large number of linked SLVs and DLVs. On a UPGMA dendrogram, isolates of the ST8 and ST11 complexes (4) appeared to be related (Fig. 7), and this finding was explored by relaxing the eBURST group definition to five of seven shared alleles. Under these conditions, isolates of both clonal complexes were placed within a single group, although they formed two separate clusters, since no ST within the ST8 complex was an SLV of any ST in the ST11 complex (Fig. 8).

A dendrogram separates STs assigned to lineage 3 of *N. meningitidis* (4, 23, 30) into two major clusters of lineages representing the ST41 and ST44 clonal complexes (Fig. 9). Isolates of both ST41 and ST44 complexes are assigned as a single clonal complex by eBURST (six of seven shared alleles), and this clonal complex is the largest so far observed by MLST for any species (411 isolates; 304 STs). ST41 was assigned as the primary founder of the lineage 3 clonal complex, with 69 SLVs (79% bootstrap support), and ST44 was identified as a large subgroup founder, with 64 SLVs (57%). The eBURST diagram (Fig. 10) confirms that the lineage 3 complex is divided into two major subgroups, the founders of which (ST44 and ST41) are connected through ST303. The ST41 subgroup is the largest, consistent with the status of ST41 as the primary founder. Curiously, ST303 is observed only three times in the database and yet has a total of 35 SLVs (25 of which are not apparent in Fig. 10, as they have been preferentially assigned to either ST44 or ST41, as these are larger subgroups; see Materials and Methods). ST41 and ST44 are both SLVs of ST303, and it is possible that ST303 is the real primary founder of this complicated (and presumably relatively old) clonal complex but now is rarely encountered among contemporary isolates.

The two major subgroups of the ST41-ST44 (lineage 3) clonal complex contain a number of subgroups, and the complexity of the diagram in Fig. 10 reflects the presumably rapid diversification of this highly successful complex. This example also further illustrates that eBURST is able to reveal possible evolutionary pathways even for the largest and most complicated clonal complexes.

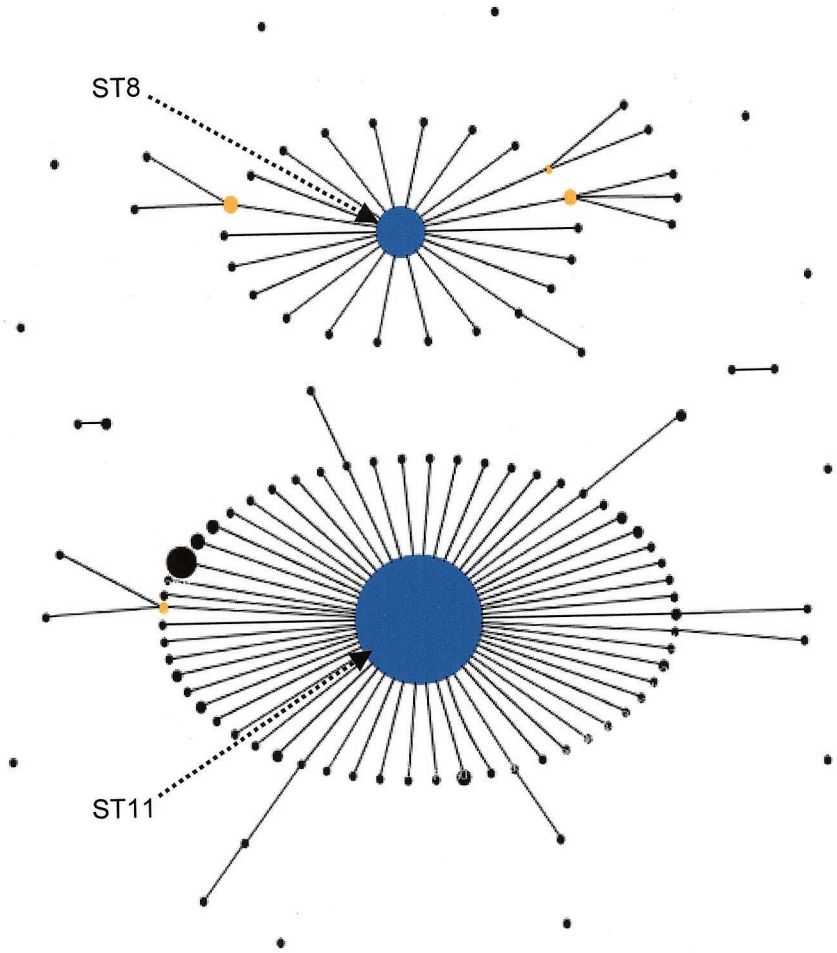

FIG. 8. Analysis of the ST8 and ST11 clonal complexes of *N. meningitidis*. eBURST groups were obtained from the entire *N. meningitidis* MLST database with the group definition of five of seven shared alleles. With this group definition, ST8 and ST11 were placed in a single group, which is displayed as an eBURST diagram. ST8 and ST11 are the primary founders of two clonal complexes (bootstrap confidence value of 100%), and most other isolates are SLVs of either ST8 or ST11; there are also two pairs of linked STs and a number of individual unlinked STs.

DISCUSSION

The relationships among isolates of bacterial species typically are displayed with a clustering algorithm, which identifies closely related genotypes but, in the absence of a realistic model of clonal expansion, provides no information about the founding genotypes or the likely patterns of evolutionary descent within the clusters. We address this important problem by using a new implementation of an algorithm that extracts this information from MLST data (or, in principle, other multilocus data). A full description of the features of eBURST is available in the documentation provided at http://eburst.mlst.net. The BURST algorithm was also recently incorporated as a set of "priority rules" into the minimum-spanning-tree method within the latest BioNumerics cluster analysis module (Applied Maths, Sint-Martens-Latem, Belgium).

The *S. pneumoniae* example is a very simple one, because the selected clonal complex is less than 50 years old and all isolates (except for a single DLV) are SLVs of the phylogenetically central ST81, which is likely the founder (35). The example is also simple because no antibiotic-susceptible isolates with genotypes similar to ST81 have been identified, and resistance appears to have occurred within a rare genotype that subsequently has increased greatly in frequency under strong selection.

S. aureus clones diversify mainly by point mutation (16), and in most cases, the clonal complexes also have a simple structure, with a single founder and a number of linked SLVs (Fig. 3); the ST239 complex is more complicated and is discussed further below.

However, the clonal complexes containing the major MRSA clones are more complex than the pneumococcal example, because resistance to methicillin has emerged in successful MSSA clones within preexisting and presumably relatively old methicillin-susceptible clonal complexes (11). In contrast, the pneumococcal Spain^{23F}-1 antibiotic-resistant ST81 clone has an allelic profile that has not been observed among antibiotic-susceptible isolates.

Clonal complexes of *C. jejuni* and *N. meningitidis* were selected for analysis because recombination rates are known to be high in these species (13, 31), resulting in rapid diversification of clones and thus providing a challenging test of the utility of the eBURST algorithm. In both species, the complex and somewhat arbitrary branching patterns among STs pro-

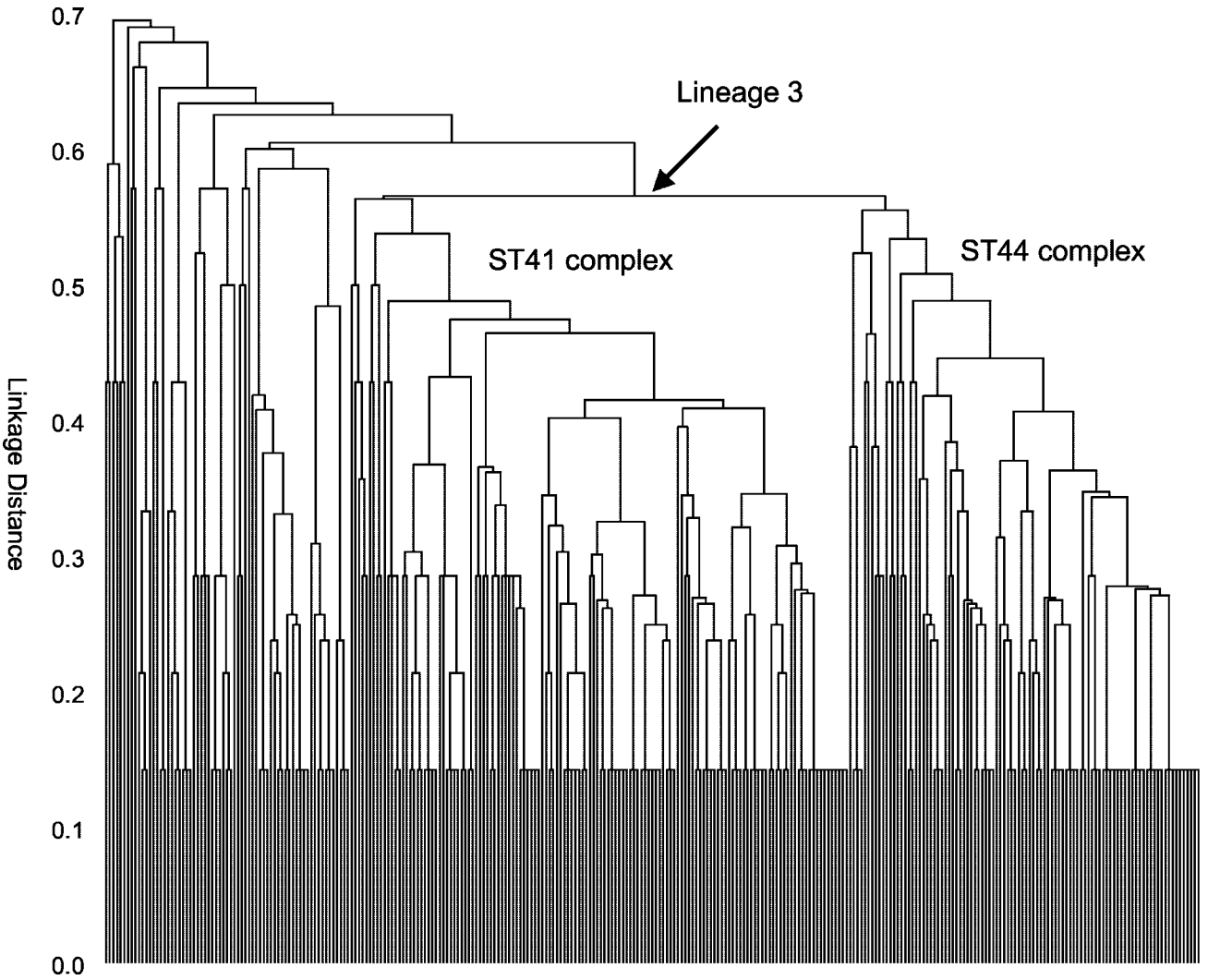

FIG. 9. Relatedness of STs of lineage 3 displayed as a dendrogram. STs that shared alleles at ≥3 of the 7 MLST loci with ST41 or ST44 were obtained from the *Neisseria* MLST website, and a dendrogram was constructed. STs assigned to lineage 3 descended from the node marked by an arrow. A major subdivision of lineage 3 into a cluster of STs that included ST41 and another that included ST44 is shown. Only one example of each ST was used in the analysis.

duced by a dendrogram were transformed by eBURST into patterns of evolutionary descent that are relatively easy to interpret. With the exception of lineage 3, which was more complicated, eBURST resolved the major meningococcal clonal complexes into a single primary founder surrounded by a large number of descendant SLVs and occasional subgroups (Fig. 6 and 8).

The advantages of the conservative approach used by eBURST, in which links are shown only between STs that differ at a single locus, are demonstrated by the analysis of the meningococcal clonal complexes. With the default group definition, eBURST shows that the great majority of isolates of both the ST8 and the ST11 clonal complexes are SLVs of their respective strongly supported primary founders. Relaxing the stringency of the group definition to five of seven shared alleles places both of these clonal complexes into a single group, although the ST8 and ST11 clusters themselves are not linked. A less conservative approach that would allow links to be drawn between DLVs would connect these two clusters, but the validity of the links would be doubtful, as links between DLVs are expected to be less robust than those between SLVs. These two clonal complexes clearly are related (5, 23) and probably emerged as two subgroups of the same clonal complex, although the precise evolutionary events that resulted in their divergence cannot be unambiguously reconstructed.

This conservative approach will result in the exclusion of some STs that should be connected to a primary founder by virtue of descent. For example, there is a single DLV of the *S. pneumoniae* Spain^{23F}-1 clone (ST81) that is multiply antibiotic resistant and therefore almost certainly descended from ST81; however, in the absence of an intermediate SLV in the MLST database, it is not linked to the other STs in the cluster. However, the benefits of the conservative approach, which attempts to identify clusters of STs with the highest level of confidence in their common descent, are considered to outweigh the omission of the occasional DLV from the eBURST diagram.

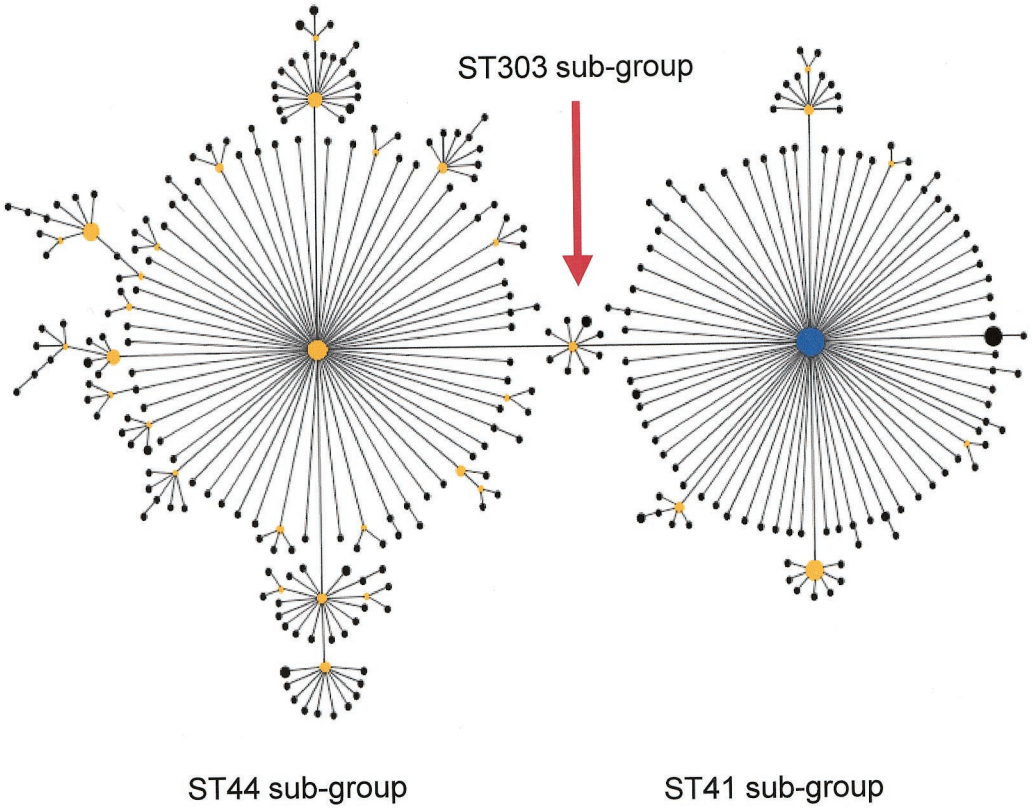

FIG. 10. Analysis of lineage 3 of *N. meningitidis*. The entire *N. meningitidis* MLST database was analyzed with the stringent (default) group definition; the group that included ST41 and ST44 is displayed as an eBURST diagram. The two main subgroups and the linking ST303 subgroup are shown. Bootstrap support values for ST41 and ST44 as the primary founders were 79 and 57%, respectively.

An eBURST diagram clearly provides far more information about founding genotypes and patterns of evolutionary descent than a dendrogram. The analysis of the major *C. jejuni* ST21 clonal complex shows that eBURST also sheds more light on its diversification than could be achieved with splits decomposition (7). eBURST confirmed ST21 as the ancestor of the *C. jejuni* ST21 clonal complex and also revealed that several SLVs of ST21 appear to have diversified to form major subgroups. It may be illuminating to map to the eBURST diagram other information about these isolates, such as host preference, although such an analysis is outside the scope of this study.

It should be stressed that the eBURST program provides only a hypothesis about the origins and patterns of descent within clonal complexes. The assignments of primary founders are likely to be correct when only a single ST in a clonal complex has very strong bootstrap support, but care should be used in inferring patterns of descent when more than one ST has considerable bootstrap support or when no ST has strong bootstrap support (which is often the case for very small complexes). It must also be emphasized that the bootstrapping procedure is designed for use with the default group definition, in which all STs are part of a single clonal complex.

The presence of two (or more) STs with good bootstrap support occurs mainly within large clonal complexes and provides an alert that the assignment of the primary founder by eBURST is unlikely to be robust and that further exploration of the data is required. The ST239 complex of *S. aureus* was used to illustrate this type of situation. Three STs within this complex are very prevalent in the *S. aureus* MLST database, and two of them have high and approximately equal bootstrap values. One of them (ST239) is predicted to be the group ancestor, as it has one more SLV, but consideration of the presence of the *mecA* gene (which confers resistance to methicillin), the structure of the *mec* region, and the frequencies of variant alleles within SLVs suggest that this prediction is incorrect and that ST8 is the most likely primary founder (11). This latter reassignment is biologically plausible, as it makes the primary founder phylogenetically central; the other two major STs become the founders of major subgroups, which are derived from ST8 by a change at a single locus (Fig. 3).

A similar situation occurs with lineage 3 of *N. meningitidis*, in which two major STs, which are DLVs of each other, have substantial bootstrap support. Although one of these is assigned as the primary founder, by analogy with the *S. aureus* ST239 complex, it is equally possible that these two STs are the founders of large subgroups and that the primary founder is the phylogenetically central ST303 (Fig. 10). On the basis of this hypothesis, ST41 and ST44 are both successful SLVs of ST303 that have diversified to become the founders of large subgroups. This example demonstrates the power of this approach, as the rarity of ST303, combined with the relatively complicated structure of the lineage 3 clonal complex, makes it very unlikely that this possible pattern of descent would have been revealed by other clustering techniques.

Even in situations in which the primary founder of a large group cannot be assigned unambiguously, the relationships between STs are still likely to approximate the true patterns of descent, and it is only the direction of descent between the different subgroups (i.e., the assignment of primary as opposed to subgroup founders) that tends to be uncertain. The problem of assigning a clear primary founder in some groups may result from a shift in ST frequencies over time, so that for old clonal complexes, there may be few examples of the primary founder (and its SLVs) relative to subgroup founders in contemporary samples of the population. This problem also may be exacerbated by sampling bias. For the ST239 clonal complex, sampling bias could have arisen from an overrepresentation of antibiotic-resistant (MRSA) strains within the data set, as these strains are of particular clinical relevance, and many strains within the *S. aureus* MLST database originate from hospital collections (9, 11).

Alternatively, natural selection may impose a bias within the population owing to the emergence of strains with a strong adaptive advantage, such as antibiotic resistance. For example, in Spain about 40% of pneumococci from carriage and disease cases are antibiotic resistant (17). A well-sampled contemporary collection of isolates from this country will be very different from that obtained 50 years ago, due to the strong selective advantage of genotypes that have become resistant, and such a major shift in ST frequencies could have an impact on the assignment of primary founders. The selective advantage of resistant strains may have led to the increase in the frequency of MRSA clone ST239 in the population, with subsequent diversification resulting in a larger number of SLVs of ST239 than of its immediate ancestor, ST8. It is not clear how to solve these problems within the confines of an algorithm, although the bootstrapping procedure can help to identify cases in which assignments of founders are not secure. When bootstrapping indicates that there may be more than one candidate primary founder, sampling bias within the data set should be considered, and any additional phenotypic, genotypic, or epidemiological data that are available should be used to examine the relative plausibility of the alternative founders and patterns of descent.

A general feature of bacterial clonal complexes is that the primary founder predicted by eBURST usually corresponds to a prevalent ST. STs that become founders of major clonal complexes (or subgroups) must predate their descendants and will have increased in frequency in the population. Thus, in the absence of strong selection and with a reasonably unbiased sampling frame, they are likely to outnumber their descendants. Examination of the eBURST diagrams shows that the primary and subgroup founders typically are prevalent STs. The number of isolates of each ST is not used by eBURST for the assignment of founders, but the predominance of STs assigned as predicted founders provides additional independent support for the assignments.

Although the assignments of primary founders, the computation of the confidence of these assignments, and the patterns of descent are all designed for use with individual clonal complexes, eBURST also can be used to produce an overall view of a bacterial population (the population snapshot). Figures 2 and 3 show examples of this type of display, which allows the overall structure of a bacterial population to be visualized. eBURST also can help to describe the clonal structures of populations in a quantitative way. For example, the number of clonal complexes observed within a population and the numbers of founders and subgroup founders (i.e., the number of nodes within a complex) provide a means of describing and comparing the structures of different populations on a purely quantitative level. However, any comparisons between populations require similar sampling frames to produce meaningful results. Finally, the identification of well-supported founding genotypes and their respective SLVs allows an estimate of the relative contributions of recombination and point mutation toward clonal diversification, as discussed elsewhere (13, 15).

ACKNOWLEDGMENTS

This work depended on the availability of the public MLST databases, which are kept in the laboratory of Brian Spratt at Imperial College London (*S. aureus* and *S. pneumoniae*, curated by Mark Enright and Angela Brueggemann) and the laboratory of Martin Maiden at the University of Oxford (*C. jejuni* and *N. meningitidis*, curated by Kate Dingle and Keith Jolley). We acknowledge all those who submit strains to these databases. We also thank Christophe Fraser for helpful discussions.

This research was supported by the Wellcome Trust. B.G.S. is a Wellcome Trust principal research fellow. E.J.F. is supported by an MRC career development award.

REFERENCES

1. **Achtman, M., K. Zurth, G. Morelli, G. Torrea, A. Guiyoule, and E. Carniel.** 1999. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. Proc. Natl. Acad. Sci. USA **96:**14043–14048.
2. **Bougnoux, M. E., S. Morand, and C. d'Enfert.** 2002. Usefulness of multilocus sequence typing for characterization of clinical isolates of *Candida albicans*. J. Clin. Microbiol. **40:**1290–1297.
3. **Caugant, D. A., L. F. Mocca, C. E. Frasch, L. O. Froholm, W. D. Zollinger, and R. K. Selander.** 1987. Genetic structure of *Neisseria meningitidis* populations in relation to serogroup, serotype, and outer membrane protein pattern. J. Bacteriol. **169:**2781–2792.
4. **Caugant, D. A.** 1998. Population genetics and molecular epidemiology of *Neisseria meningitidis*. APMIS **106:**505–525.
5. **Claus, H., H. Weinand, M. Frosch, and U. Vogel.** 2003. Identification of the hypervirulent lineages of *Neisseria meningitidis*, the ST-8 and ST-11 complexes, by using monoclonal antibodies specific to *Nme*DI. J. Clin. Microbiol. **41:**3873–3876.
6. **Crisostomo, M. I., H. Westh, A. Tomasz, M. Chung, D. C. Oliveira, and H. de Lencastre.** 2001. The evolution of methicillin resistance in *Staphylococcus aureus*: similarity of genetic backgrounds in historically early methicillin-susceptible and -resistant isolates and contemporary epidemic clones. Proc. Natl. Acad. Sci. USA **98:**9865–9870.
7. **Dingle, K. E., F. M. Colles, D. R. A. Wareing, R. Ure, A. J. Fox, F. E. Bolton, H. J. Bootsma, R. J. L. Willems, R. Urwin, and M. C. J. Maiden.** 2001. Multilocus sequence typing system for *Campylobacter jejuni*. J. Clin. Microbiol. **39:**14–23.
8. **Enright, M. C., and B. G. Spratt.** 1998. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology **144:**3049–3060.
9. **Enright, M. C., N. P. J. Day, C. E. Davies, S. J. Peacock, and B. G. Spratt.** 2000. Multilocus sequence typing for the characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. J. Clin. Microbiol. **38:**1008–1015.
10. **Enright, M. C., B. G. Spratt, A. Kalia, J. H. Cross, and D. E. Bessen.** 2001. Multilocus sequence typing of *Streptococcus pyogenes* and the relationships between *emm* type and clone. Infect. Immun. **69:**2416–2427.
11. **Enright, M. C., D. A. Robinson, G. Randle, E. J. Feil, H. Grundmann, and B. G. Spratt.** 2002. Evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). Proc. Natl. Acad. Sci. USA **99:**7687–7692.
12. **Falush, D., T. Wirth, B. Linz, J. K. Pritchard, M. Stephens, M. Kidd, M. J. Blaser, D. Y. Graham, S. Vacher, G. I. Perez-Perez, Y. Yamaoka, F. Megraud, K. Otto, U. Reichard, E. Katzowitsch, X. Wang, M. Achtman, and S. Suerbaum.** 2003. Traces of human migrations in *Helicobacter pylori* populations. Science **299:**1582–1585.
13. **Feil, E. J., M. C. J. Maiden, M. Achtman, and B. G. Spratt.** 1999. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. Mol. Biol. Evol. **16:**1496–1502.
14. **Feil, E. J., E. C. Holmes, D. E. Bessen, M.-S. Chan, N. P. J. Day, M. C.**

Enright, R. Goldstein, D. W. Hood, A. Kalia, C. E. Moore, J. Zhou, and B. G. Spratt. 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. USA **98:**182–187.

15. **Feil, E. J., and B. G. Spratt.** 2001. Recombination and the population structures of bacterial pathogens. Annu. Rev. Microbiol. **55:**561–590.
16. **Feil, E. J., J. E. Cooper, H. Grundmann, D. A. Robinson, M. C. Enright, A. Berendt, S. Peacock, J. Maynard Smith, M. Murphy, B. G. Spratt, C. E. Moore, and N. P. J. Day.** 2003. How clonal is *Staphylococcus aureus*? J. Bacteriol. **185:**3307–3316.
17. **Fenoll, A., I. Jado, D. Vicioso, A. Perez, and J. Casal.** 1998. Evolution of *Streptococcus pneumoniae* serotypes and antibiotic resistance in Spain: update (1990 to 1996). J. Clin. Microbiol. **36:**3447–3454.
18. **Godoy, D., G. Randle, A. J. Simpson, D. Aanensen, T. L. Pitt, R. Kinoshita, and B. G. Spratt.** 2003. Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. J. Clin. Microbiol. **41:**2068–2079.
19. **Homan, W. L., D. Tribe, S. Poznanski, M. Li, G. Hogg, E. Spalburg, J. D. Van Embden, and R. J. Willems.** 2002. Multilocus sequence typing scheme for *Enterococcus faecium*. J. Clin. Microbiol. **40:**1963–1971.
20. **Jolley, K. A., J. Kalmusova, E. J. Feil, S. Gupta, M. Musilek, P. Kriz, and M. C. J. Maiden.** 2000. Carried meningococci in the Czech Republic: a diverse recombining population. J. Clin. Microbiol. **38:**4492–4498.
21. **Jones, N., J. F. Bohnsack, S. Takahashi, K. A. Oliver, M.-S. Chan, F. Kunst, P. Glaser, C. Rusniok, D. W. Crook, R. M. Harding, N. Bisharat, and B. G. Spratt.** 2003. Multilocus sequence typing system for group B streptococcus. J. Clin. Microbiol. **41:**2530–2536.
22. **Kidgell, C., U. Reichard, J. Wain, B. Linz, M. Torpdahl, G. Dougan, and M. Achtman.** 2002. *Salmonella typhi*, the causative agent of typhoid fever, is approximately 50,000 years old. Infect. Genet. Evol. **2:**39–45.
23. **Maiden, M. C. J., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. USA **95:**3140–3145.
24. **McGee, L., L. McDougal, J. Zhou, B. G. Spratt, F. C. Tenover, R. George, R. Hakenbeck, W. Hryniewicz, J.-C. Lefévre, A. Tomasz, and K. P. Klugman.** 2001. Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the Pneumococcal Molecular Epidemiology Network (PMEN). J. Clin. Microbiol. **39:**2565–2571.
25. **Meats, E., E. J. Feil, S. Stringer, A. J. Cody, R. Goldstein, J. S. Kroll, T. Popovic, and B. G. Spratt.** 2003. Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. J. Clin. Microbiol. **41:**1623–1636.
26. **Musser, J. M., J. S. Kroll, E. R. Moxon, and R. K. Selander.** 1988. Clonal population structure of encapsulated *Haemophilus influenzae*. Infect. Immun. **56:**1837–1845.
27. **Musser, J. M., and V. Kapur.** 1992. Clonal analysis of methicillin-resistant *Staphylococcus aureus* strains from intercontinental sources: association of the *mec* gene with divergent phylogenetic lineages implies dissemination by horizontal transfer and recombination. J. Clin. Microbiol. **30:**2058–2063.
28. **Reid, S. D., C. J. Herbelin, A. C. Bumbaugh, R. K. Selander, and T. S. Whittam.** 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature **406:**64–67.
29. **Salcedo, C., L. Arreaza, B. Alcala, L. de la Fuente, and J. A. Vazquez.** 2003. Development of a multilocus sequence typing method for analysis of *Listeria monocytogenes* clones. J. Clin. Microbiol. **41:**757–762.
30. **Scholten, R. J., J. T. Poolman, H. A. Valkenburg, H. A. Bijlmer, J. Dankert, and D. A. Caugant.** 1994. Phenotypic and genotypic changes in a new clone complex of *Neisseria meningitidis* causing disease in The Netherlands, 1958–1990. J. Infect. Dis. **169:**673–676.
31. **Schouls, L. M., S. Reulen, B. Duim, J. A. Wagenaar, R. J. Willems, K. E. Dingle, F. M. Colles, and J. D. Van Embden.** 2003. Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. J. Clin. Microbiol. **41:**15–26.
32. **Selander, R. K., J. M. Musser, D. A. Caugant, M. N. Gilmour, and T. S. Whittam.** 1987. Population genetics of pathogenic bacteria. Microb. Pathog. **3:**1–7.
33. **Spratt, B. G.** 1999. Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the Internet. Curr. Opin. Microbiol. **2:**312–316.
34. **Wareing, D. R. A., R. Ure, F. M. Colles, F. J. Bolton, A. J. Fox, M. C. J. Maiden, and K. E. Dingle.** 2003. Reference isolates for the clonal complexes of *Campylobacter jejuni*. Lett. Appl. Microbiol. **36:**106–110.
35. **Zhou, J., M. C. Enright, and B. G. Spratt.** 2000. Identification of the major Spanish clones of penicillin-resistant pneumococci via the Internet using multilocus sequence typing. J. Clin. Microbiol. **38:**977–986.